International Journal of Engineering and Applied Computer Science (IJEACS)
Volume: 02, Issue: 06, ISBN: 978-0-9957075-7-3, June 2017

Implementation of Multi-node Clusters in Column Oriented Database using HDFS

P. Naresh
SICET, Hyderabad, India

G. Nataraja Shekhar
SICET, Hyderabad, India

M. Kiran Kumar
SICET, Hyderabad, India

P. Rajyalakshmi
HITS Hyderabad, India

***Abstract*—Generally HBASE is NoSQL database which runs in the Hadoop environment, so it can be called as Hadoop Database. By using Hadoop distributed file system and map reduce with the implementation of key/value store as real time data access combines the deep capabilities and efficiency of map reduce. Basically testing is done by using single node clustering which improved the performance of query when compared to SQL, even though performance is enhanced, the data retrieval becomes complicated as there is no multi node clusters and totally based on SQL queries. In this paper, we use the concepts of HBase, which is a column oriented database and it is on the top of HDFS (Hadoop distributed file system) along with multi node clustering which increases the performance. HBase is key/value store which is Consistent, Distributed, Multidimensional and Sorted map. Data storage in HBase in the form of cells, and here those cells are grouped by a row key. Hence our proposal yields better results regarding query performance and data retrieval compared to existing approaches.**

Keywords- Database; Cluster; Hadoop; HDFS; MapReduce.

I. INTRODUCTION

The term "Big Data" denotes to the unremitting huge expansion in the volume and variety along with velocity of data processing. Volume denotes to the scale of the data and processing needs. Variety denotes to the diverse layouts of information like images, booklets, videos, streams. Velocity denotes to the rapidity at which the information is being produced. Data is increasing extremely by day to day. The IDC evaluation and put the dimension of the "digital universe" at 0.18 zettabytes in 2006, and is projecting a tenfold growth by 2011 to 1.8 zettabytes.

- Structured data: It describes data that resides in fixed field files record tables, and it is standard database and data contained in relational databases and spreadsheets. The data arrangement and consistency permits it to react to basic quires to land at usable data in view of association's parameters.
- Semi-structured data: It is a database model where there is no division between the data and schema, it is not a permanent schema and it contains labels or different markers to uphold chains of importance of records and fields inside of the information.
- Unstructured data: It alludes to data that either does not have a predefined information show or is not sorted out in pre-characterized way. This sort of information can't without much of a stretch be recorded into social tables for investigation or questioning like pictures, sound, and feature documents.
- Velocity: Commonly velocity refers to the speed of the data processing. Ordinary comprehension of speed regularly is how rapidly the information arrives and put away and how rapidly, it can be retrieved. Velocity ought to additionally be connected to information in movement. The rate at which the information is streaming is found. The different data streams and the increment in sensor system arrangement have prompted a consistent stream of information at a pace that has made it unimaginable for customary frameworks to handle.

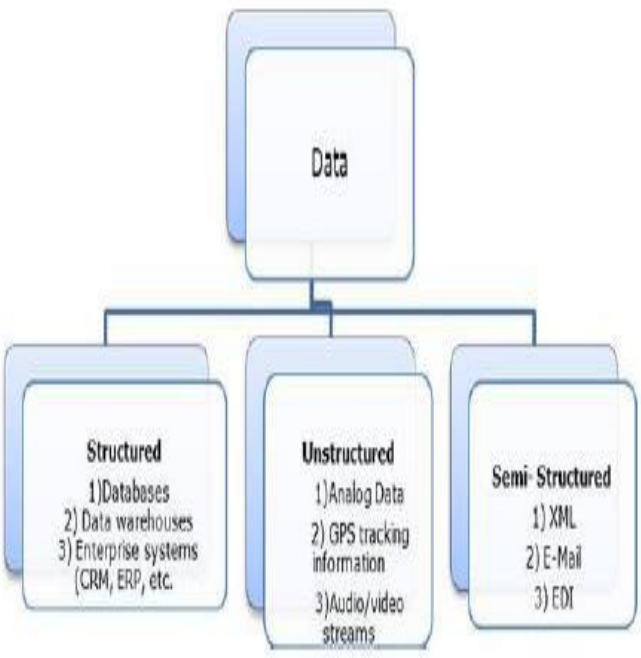

Figure 1. Types of data.

MapReduce is a programming model for expansive scale data preparing and MapReduce [5]. Programs written in different dialects like Java, Python, C++, Ruby, and so forth one essential point is the MapReduce projects are naturally parallel. MapReduce carries its own weight for substantial datasets. Here Datasets is nothing but a data can be stored as a semi-structured and record-oriented and data arrangement bolsters a rich arrangement of meteorological components, a large number of which are discretionary or with variable information lengths.

- Restricted parallel programming model meant for large clusters.
- Map Reduce divides the workload into multiple independent tasks and schedule them across cluster nodes.
- In a Map Reduce cluster [4], data is distributed to every one of the hubs of the group as it is being stacked in.

HDFS refers to "Hadoop Distributed File System" [2] HDFS is designed for the storing of large documents with gushing information access examples, running on bunches of product equipment. The key part of HDFS is its architecture itself.

A. *HDFS Architecture*

- The HDFS is designed in the form of a master and slave pattern.
- Here, the master is the NameNode and the slave is the DataNode.

B. *NameNode*

Here NameNode acts as a master node of HDFS and it organize all the data node's file system operations and it maintain replication node and maintains the file system tree which stored at local disk in the form of two files.

C. *NameSpace Image*

- The purpose of this is that it captures the snapshot of what the file system comprises of.
- The below is how a namespace image would be in the name node.
- This consists of the file_name, replication factor and block_ids.

We know that enormous data turns out consistently from the web. In previous this huge area of data is taken care of by Relational database administration frameworks (RDBMS) [4]. But know a day's predominated technology is there so at a time RDBMS can't handle by structured and semi-structured and unstructured data. The new technology NoSql is a non-relational database administration framework, not the same as customary relational database administration frameworks in some critical ways. It is intended for distributed data stores where huge size of information putting away need.

II. RELATED WORK

WUMUTI·NAHEMAN, IANXINWEI [1] they introduce the structural planning and information model of HBase database and is illustrative of NoSQL databases and did some performance test on HBase database include column family test according to their results he says query rate of HBase is moderate under single machine environment , however can be altogether enhanced in multi-machine bunch environment. Furthermore, they know the network data it is big challenge in the market it is explosive growth and NoSQL databases [2] have been widely used in some scenarios it can be combination of relational database to make up the defects of their own and above results of NoSQL database are not mature enough so his future research is integration of relational database and NoSQL database.

Jyoti Nandimath, Ankur Patil, Ekata Banerjee, Pratima Kakade, Saumitra Vaidya [7] confronted issues of effectiveness, execution and lifted base expense with the information handling in the unified environment with help of conveyed structural engineering so these expansive associations had the capacity defeat the issues of removing applicable data from a colossal information dump. This is the one of the best devices utilized as a part of the information handling issues in Apache hadoop, using the Apache Hadoop's different segments, for example, an information groups, guide decrease calculations and circulated preparing.

A. *Row-Oriented Data Stores*

- Information is put away and recovered one column at once and subsequently could read pointless information if a percentage of the information consecutively is needed.
- Easy to peruse and compose record
- Well suited for OLTP framework
- Not effective in performing operations pertinent to the whole dataset and subsequently conglomeration is an extravagant operation.
- Regular pressure instruments give less successful results than those on segment situated information store.

B. *Column-Oriented Data Stores*

- Information is put away and recovered in segments and subsequently can read just important information if some information is needed.
- Perused and Write are commonly slower operations.
- Appropriate for OLAP frameworks.
- Can effectively perform operations pertinent to the whole dataset and thus empowers total over numerous lines and sections.
- Allows high pressure rates because of couple of particular qualities in segments.

III. IMPLEMENTATION

HBase is a column oriented database that executed by Google's Big Table stockpiling building design. It can oversee organized and semi-organized information and has some implicit elements, for example, versatility, forming, pressure and waste accumulation. Since its uses compose ahead logging and disseminated setup, it can give adaptation to internal failure and snappy recuperation from individual server disappointments.

HBase based on Hadoop/HDFS [8] and the information put away in HBase can be controlled utilizing Hadoop's MapReduce[5] abilities. How about we now investigate how HBase (a segment situated database) is unique in relation to some other information structures and ideas that we are acquainted with Row-Oriented versus Segment Oriented information stores. As demonstrated as follows, consecutively arranged information store, a line is a unit of information that is perused or composed together. In a section arranged information store, the information in a segment is put away together and subsequently immediately recovered.

Row ID	Customer	Product	Amount
101	John White	Chairs	$400.00
102	Jane Brown	Lamps	$500.00
103	Bill Green	Lamps	$150.00
104	Jack Black	Desk	$700.00
105	Jane Brown	Desk	$650.00
106	Bill Green	Desk	$900.00

Figure 2. Simple HBase Table.

The Data Model in HBase is intended to suit semi-organized information that could differ in field size, information sort and segments. Moreover, the format of the information model makes it less demanding to segment the information and appropriate it over the bunch. The Data Model in HBase is made of diverse consistent parts, for example, Tables, Rows, Column Families, Columns, Cells and Versions.

A column is one example of information in a table and is distinguished by a rowkey. Rowkeys are one of a kind in a Table and are constantly regarded as a byte. All lines are dependably lexicographically by their column key. Example:

Hbase (main):001:0> scan 'table1'

ROW COLUMN+CELL

row1 column=cf-1:, timestamp=1297073325971 ..

row10 column=cf-1:, timestamp=1297073337383 ...

row11 column=cf-1:, timestamp=1297073340493 ...

row-2 column=cf1:, timestamp=1297073329851 ...

row-22 column=cf1:, timestamp=1297073344482 ...

row-3 column=cf1:, timestamp=1297073333504 ...

row-abc column=cf1:, timestamp=1297073349875 ...

7 row(s) in 0.1100 seconds

Observe very carefully the numbering is not sequence because of random sorting order. In lexicographical sorting, and every key is compared through binary level, byte by byte from left to right.

A. *Column Families*

Information consecutively is gathered together as Column Families. Every Column Family has one more Columns and these Columns in a family are put away together in a low level stockpiling record known as HFile. Segment Families frame the essential unit of physical stockpiling to which certain HBase.

The table beneath shows Customer and Sales Column Families. The Customer Column Family is made up 2 sections – Name and City, though the Sales Column Families is made up to 2 segments – Product and Am.

Row Key	Customer		Sales	
Customer Id	Name	City	Product	Amount
101	John White	Los Angeles, CA	Chairs	$400.00
102	Jane Brown	Atlanta, GA	Lamps	$200.00
103	Bill Green	Pittsburgh, PA	Desk	$500.00
104	Jack Black	St. Louis, MO	Bed	$1600.00

Column Families

Figure 3. Column Family.

IV. SYSTEM ANALYSIS

Generally cluster means group of similar things or occurring closely together. A Hadoop bunch is an exceptional sort of computational group planned particularly for putting away and examining gigantic measures of unstructured information in appropriated processing environment. Hadoop bunch is open source disseminated preparing programming keep running on thing PCs.

In Hadoop cluster environment, we have many machines are there but here we choose the following way one machine in the bunch assigned as the Name Node and another machine go about as a Job Tracker, these Name Node and Job Tracker are the masters and rest of the machine act as both Data Node and Task Tracker these two are the slaves. And here we remember that have a network connection between nodes. Hadoop clusters are used for boosting the pace of information investigation applications. Hadoop clusters are high resistance to failure due to every bit of the information is replicated onto other bunch hubs so information is not lose if hub falls flat.

Single hub bunch can be valuable to work application focus on a solitary server and Application focus treats a solitary hub group or stand-alone server single-hub bunches can profit by application focus without working in bunch environment. Before Installation of java it can require a working java 1.5 installations and it follow the commands like

$sudo apt-get update

$sudo apt-get install sun-java6-jdk

$sudo update-java-alternatives -s java-6-sun

Cluster run in only one machine and it did not take multiple machines but in this single node machine we have to create the tables and column families and this data can be stored in the HDFS database and tables are splits into regions and this regions are served by the region servers and here query performance is slow because of single cluster and see the results on single node cluster one by one. We already know the HBase [6] is a section arranged database it is open source usage of the BigTable and it is based on top of the Hadoop and HDFS and the information put away in HBase can be embarrassed utilizing Hadoop's MapReduce abilities. It is acquainted with Row-Oriented and Column-Oriented information stores and it can be intended to store Denormalized information contains wide and inadequately populated tables and bolster Automatic Partitioning. It is assembled for low Latency operations and gives the entrance to single lines from billions of records and information is gotten to through shell commands.

$start-hbase.sh

$hbase shell

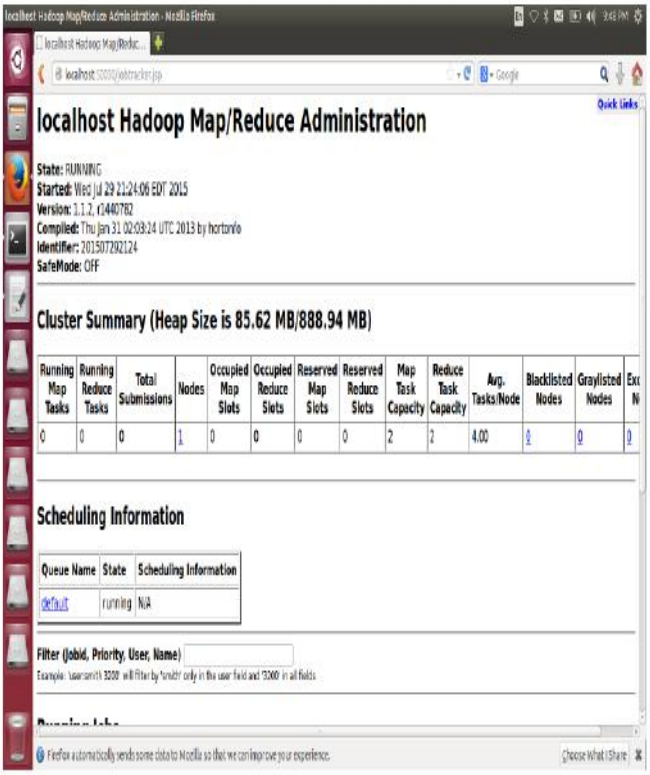

Figure 4. Map-Reduce Capabilities.

Figure 5. Multi-node cluster set-up health check-up.

V. CONCLUSION

We know the world become a narrower and present generation peoples are expect the reaction of the meticulous products, actions, issues rapid on the web so we use the automatic sentiment analysis it can be recognize and foresee present and future patterns, item audits individuals verdict about social issues. Observe the present market, the incident about individuals data will be generated day-day very large-scale so this data will hoard the concepts of Google, the concepts like Hadoop and their eco-system here ecosystems are developed by some Google developers in that eco-system provide HBase so we are focussed on performance the multi-node cluster unit by using Hbase and HDFS.

REFERENCES

[1] WUMUTI. NAHEMAN, JIANXIN WEI, "Review of NoSQL Databases and performance Testing on HBase", International Conference on Mechatronic Sciences Electric Engineering and Computer (MEC), Dec 20-22, 2013, Shenyang, China.

[2] Lizhi, Shidong Huang, Leilei Chen, Yang Zheng, "Performance Analysis and Testing ofHBase Based on Its Architecture", International Conference on IEEE Machatronic Science and ISBN:978-1-4799-0174-6/13.

[3] A B M Moniruzzaman and Syed Akhter Hossain, "NoSQL Database: New Era of Databases for Big data Analytics Classification, Characteristics and Comparison", International Journal of Database Theory and Application vol. 6, No. 4. 2013.

[4] D. Carstoiu, A. Cernian, A. Olteanu, "Hadoop HBase-0.20.2 Performance Evaluation", "Politehnica" University of Bucharest. IEEE Journal Paper.

[5] I. Tomasic, A. Rashkovska and M. Depolli. "Using Hadoop MapReduce in a Multicluster Environment", MIPRO 2013, May 20-24, 2013, Opatija, Croatia.

[6] Zhou Quan, Chunming Tang, "TSHC: Trusted Scheme for Hadoop Cluster", 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies.

[7] JyothiNndimath, AnkurPatil, "BigbData Analysis Using Apache Hadoop" IEEE IRI 2013, August 14-16, 2013 IEEE, San Francisco, California, USA.

[8] Ognjen V. Joldzic, Dijana R. Vukovic, "The Impact of Cluster Characteristics on HiveQL Query Optimization", 21st Telecommunications forum TELFOR 2013.

[9] Mehmet C.Okur, Mustafa Buyukkececi, "Big Data Challenges in Information Engineering Curriculum", 2014 IEEE International Journal Paper.

[10] Santhosh Kumar Gajendran, "A Survey on NoSQL Databases ", IEEE NoSQL Database survey paper.